Reducing OCR errors by combining two OCR systems

Reducing OCR Errors by Combining Two OCR Systems.

Martin Volk, Torsten Marek and Rico Sennrich

Institute of Computational Linguistics
University of Zurich

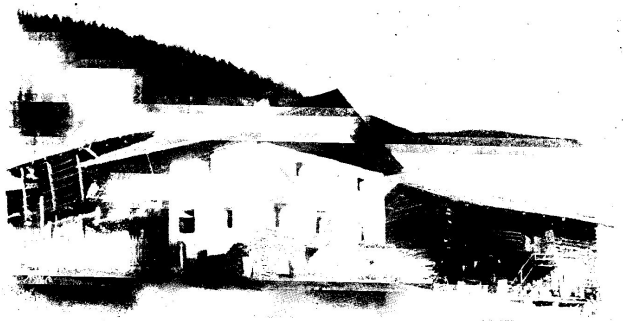
16 August 2010

The Text+Berg Project

- Corpus of "Alpine" texts
- Mountaineering and travel reports in multiple languages
- For studies on language development, culture, technology, ...
- For geo-tagging, machine translation, ...
- First step: Publications of the Swiss Alpine Club since 1864

streifen, der das ganze Thal überzieht, steht in gleicher Meereshöhe eine ständig bewohnte Hütte, worin eine alte, tapfere Jungfrau mit ihren Ziegen, Katzen und Hühnern haushaltet; und willig trotzt sie allen Winterstürmen und lauscht, ob der Frühling wiederkehre, der sie hervorlocke zur Bestellung ihres Äckerleins. Keine besondere Verlassenheit, kein Mangel zwingt die resolute Person zu ihrer Einsamkeit, und wer sie fragt, ob sie sich darin nicht unglücklich fühle, wird nicht übel heimgeschickt. So viel und so wenig braucht der Mensch manchmal, um zufrieden zu sein!

Nur wenige menschliche Wohnungen steigen in Davos noch höher ins Gebirge, und locken zum Ausharren auch im Winter in einer Region, die keinen Ackerbau mehr kennt, und nur der Viehzucht und einem Gemüsegärtchen noch eine Stätte gewährt. Auf der Schatzalp ob Davosplatz,



Bauernhaus in der Kumme.
Photographie von *A. Blum* in Davos.

an der obern Waldgrenze (1875 m), hat der Fremdenbesuch auch im Winter zu ständiger Besiedelung geführt; im Flüelathal ist es die Paßstraße, die in dem wilden Hochgebirgsthal noch auf der „Alpenrose" bei 1830 Metern und auf dem Tschuggen bei 1941 Metern den Menschen dem rauhen Klima trotzen läßt. Das Flüelahospiz, den größten Teil des Jahres hindurch in arktischer Umgebung (2388 m), liegt schon auf Engadinerboden. Das Dischmathal, an den Gletschern der Scaletta endigend, öde und bald des Waldes entbehrend, ist zur Winterszeit nur im vordern Thal bewohnt, und die letzten, ständig besetzten Häuser stehen schon bei 1700 Metern. Viel spärlicher bewohnt als das freundliche Sertig, kennen Flüela und Dischma kaum mehr den Ackerbau.

Schon zeitig im Frühling, viel eher, als wohl der Tiefländer glaubt, der bis in den Sommer die Alpengegenden noch unter hohen Schnee-

Sicherheit, Medizin, Rettungswesen

Sicurezza, medicina, soccorso in montagna

Sécurité, médecine, sauvetage

Zecken in den Klettergärten

Die Blutsauger klettern immer höher

Ein Teil der Zecken in der Schweiz übertragen Krankheiten, unter anderem das FSME-Virus, das eine schwere Hirnhautentzündung auslösen kann. Eine Studie belegt nun, dass FSME-verseuchte Zecken auch in Höhenlagen vorkommen, in denen sie bisher nicht vermutet worden sind.

Die milden Temperaturen locken Alpinisten, Sportkletterinnen und Wanderer wieder vor die Tür. Doch die Wärme, die zum Tragen kurzer und leichter Kleidung verleitet, hat auch ihre Tücken. So lauern etwa Zecken in feuchtem Unterholz, an Waldrändern und in Hecken auf ihre Opfer. Hier haben die kleinen, teilweise lauffreudigen Zecken genügend Chancen, auf ungeschützte Haut zu gelangen. Die von ihnen bevorzugten Körperstellen sind Kniekehlen, Schamgegend, Bauchnabel, Achselhöhlen, Schultern, Nacken und hinter den Ohren.

Nicht nur Wandernde müssen aufmerksam sein. Die Gefahr, dass die Blutsauger an der Kleidung haften bleiben und sich auf die Suche nach freien Körperstellen machen, ist auch im Zustieg zu Klettergebieten gross, führen die Wege doch oft durch niedriges Gebüsch und Gras. Hinzu kommt, dass die Tiere von der Klimaerwärmung profitieren und nun neuerdings auch in Höhenlagen von bis zu 1500 Metern zu finden sind. Bislang fand man Zecken nur bis in Höhen von 1000 Metern. Auch wer sich in diesen Gebieten aufhält, läuft Gefahr, mit den Tieren in Kontakt zu kommen, meldeten die Virologen des Labors Spiez kürzlich.

Zecken in Lauerstellung, als REM-Aufnahme und nachträglich eingefärbt. Vorne deutlich zu sehen sind ihre zwei scherenartigen Mundwerkzeuge (Cheliceren). Mit diesen reisst sie die Haut des Wirtes auf und gräbt mit dem Stechrüssel (Hypostom) ein Loch in das Gewebe.

Foto: www.zecken.de

Zeckenkrankheiten

Zecken in der Schweiz können zwei Krankheiten mit schweren Folgen übertragen: FSME (Frühsommer-Meningoenzephalitis) und Lyme-Borreliose. Das FSME-Virus tritt meist im Frühsommer auf und wird von Zecken in Hochrisiko- und Risikogebieten übertragen. Eine Impfung ist möglich. Nach einer Ansteckung können jedoch nur noch die Symptome, nicht aber die Krankheitsursache behandelt werden. Immunität ist nur bei durchgemachter Krankheit lebenslänglich gewährleistet, die Impfung hingegen muss alle drei Jahre aufgefrischt werden.
Die sogenannte Lyme-Borreliose wird von rund einem Drittel der Zecken übertragen. Der Erreger greift Haut, Gelenke, Nervensystem, Muskeln und Herz an. Eine Impfung dagegen ist nicht möglich, jedoch eine Behandlung mit Antibiotika nach einem Biss, wobei danach keine Immunität gegen den Erreger besteht. Wertvolle Infos gibt es unter http://www.labor-spiez.ch/de/akt/index.htm, unter www.bag.admin.ch, Stichwort FSME, oder unter www.zecken.de

Suche in aufwendiger Kleinarbeit

Problematisch war bis anhin, dass niemand genau wusste, wo sich die FSME-befallenen Zecken aufhalten. FSME-Erkrankungen mussten zwar bisher ans Bundesamt für Gesundheit (BAG) gemeldet werden. Doch leider können die wenigsten sagen, woher genau sie die Zecken haben. Zudem tritt die Krankheit erst einige Tage nach dem Zeckenbiss auf. Daher konnten Ärzte oft nur vage Informationen über den Ort der Infizierung ans BAG weitergeben.

Eine Studie des Labors Spiez in Zusammenarbeit mit Spezialisten der ABC-Abwehr-Truppe der Armee liefert nun genauere Daten. Vom April bis zum Juli 2009 wurden an 165 Orten in der Schweiz 62 343 Zecken eingesammelt. «Untersucht wurden Gebiete, aus denen Krank-

University of Zurich
Institute of Computational Linguistics

- We use commercial OCR systems (Abbyy FineReader and OmniPage)
- OCR systems are language-dependent
- Challenges during OCR of the Text+Berg corpus
  - Corpus is diachronic: layout, spelling changes regularly
  - Corpus is multilingual: German, French, Italian, English, Romansch etc.
  - Swiss spelling deviates from German norm
  - Many names (toponyms) that are not in OCR dictionary

Bauernhaus in der Kumme.

Photographie von *A. Blum* in Davos.

an der obern Waldgrenze (1875 m), hat der Fremdenbesuch auch im Winter zu ständiger Besiedelung geführt; im Flüelathal ist es die Paßstraße, die in dem wilden Hochgebirgsthal noch auf der „Alpenrose“ bei 1830 Metern und auf dem Tschuggen bei 1941 Metern den Menschen dem rauhen Klima trotzen läßt. Das Flüelahospiz, den größten Teil des Jahres hindurch in arktischer Umgebung (2388 m), liegt schon auf Engadinerboden. Das Dischmathal, an den Gletschern der Scaletta endigend, öde und bald des Waldes entbehrend, ist zur Winterszeit nur im vordern Thal bewohnt, und die letzten, ständig besetzten Häuser stehen schon bei 1700 Metern. Viel spärlicher bewohnt als das freundliche Sertig, kennen Flüela und Dischma kaum mehr den Ackerbau.

University of Zurich
Institute of Computational Linguistics

Bauernhaus in der Kumme.
Photographie von *A. Bluum* in Davos.

an der obern Waldgrenze (1875 m), hat der Fremdenbesuch auch im Winter zu ständiger Besiedelung geführt; im Flüelathal ist es die Paßstraße, die in dem wilden Hochgebirgsthal noch auf der „Alpenrose“ bei 1830 Metern und auf dem Tschuggen bei 1941 Metern den Menschen dem rauhen Klima trotzen läßt. Das Flüelahospiz, den größten Teil des Jahres hindurch in arktischer Umgebung (2388 m), liegt schon auf Engadinerboden. Das Dischmathal, an den Gletschern der Scaletta endigend, öde und bald des Waldes entbehrend, ist zur Winterszeit nur im vordern Thal bewohnt, und die letzten, ständig besetzten Häuser stehen schon bei 1700 Metern. Viel spärlicher bewohnt als das freundliche Sertig, kennen Flüela und Dischma kaum mehr den Ackerbau.

University of Zurich
Institute of Computational Linguistics

Bauernhaus in der Kumme.
Photographie von *A. Blumm* in Davos.

an der obern Waldgrenze (1875 m), hat der Fremdenbesuch auch im Winter zu ständiger Besiedelung geführt; im Flüelathal ist es die Paßstraße, die in dem wilden Hochgebirgsthal noch auf der „Alpenrose“ bei 1830 Metern und auf dem Tschuggen bei 1941 Metern den Menschen dem rauhen Klima trotzen läßt. Das Flüelahospiz, den größten Teil des Jahres hindurch in arktischer Umgebung (2388 m), liegt schon auf Engadinerboden. Das Dischmathal, an den Gletschern der Scaletta endigend, öde und bald des Waldes entbehrend, ist zur Winterszeit nur im vordern Thal bewohnt, und die letzten, ständig besetzten Häuser stehen schon bei 1700 Metern. Viel spärlicher bewohnt als das freundliche Sertig, kennen Flüela und Dischma kaum mehr den Ackerbau.

Our Tuning of OCR Systems

- Compiling word lists with 19th century spelling (1500 word forms)
  - *Thal*, *Thür* instead of *Tal*, *Tür* (valley, door)
  - *passiren* instead of *passieren*
- Compiling word lists with Swiss spelling (5000 word forms)
  - ss instead of ß
- Inclusion of geographical names from Swiss Federal Office of Topography
  - 15 000 toponyms

Fully Manual OCR Correction

- Performed by project members for a few books
- Time-intensive (dozens of hours per book)
- Not perfect: errors may be overlooked

University of Zurich
Institute of Computational Linguistics

- Substitution rules
- Several substitutions for a character sequence possible:
  ii can be *n*, *u*, *ü*, *li* or *il*
- Most frequent[1] alternative (on a word level) is selected:
  iiberein → überein

[1] We currently use the latest Text+Berg release for our frequency counts

OCR System Combination

Our goal: Combination of two OCR systems that:

- requires little manual work
- works well with Text+Berg corpus.
- other approaches problematic:
  - dictionary-based approach suboptimal (toponyms / special terminology)
  - multilingual text

OCR System Combination

- Primary system: Abbyy Finereader 7 (AF7)
- Secondary system: OmniPage 17 (OP17)
- We train a unigram language model on the uncorrected Text+Berg corpus (created with AF7)
  $\longrightarrow$ bootstrapped decision procedure
- Whenever OCR output of the two systems differs, we select the more probable (i.e. more frequent) alternative.
- In case several alternatives are equally probable, we pick AF7 output

Results: Automatic Evaluation

These are not the final results!

Evaluation with manually corrected 1899 yearbook (488 pages; 220 000 tokens).

Table: Word accuracy of different OCR systems.

system	word errors	word accuracy
Abbyy FineReader 7	1260	99.26%
OmniPage 17	6466	96.21%
merged system	1305	99.24%

University of Zurich
Institute of Computational Linguistics

These are not the final results!

Evaluation with manually corrected 1899 yearbook (488 pages; 220 000 tokens).

Table: Word accuracy of different OCR systems.

system	word errors	word accuracy
Abbyy FineReader 7	1260	99.26%
OmniPage 17	6466	96.21%
merged system	1305	99.24%

Starting point for manual correction was AF7 output: evaluation is skewed in favor of AF7.

Results: Manual Evaluation

- Alignment algorithm returns 1800 differences between two systems
- In 1350 cases, AF7 alternative is chosen (trivial case)
- We manually analysed the other 450 cases:
  - 277 errors fixed
  - 82 errors added
  - 89 equally good/bad alternatives
  - **net gain: 195 improvements**, out of 1000-2000 OCR errors

Results: Examples

Table: Examples where OmniPage is preferred over FineReader by our decision procedure.

Abbyy FineReader	OmniPage	correct alternative in context	jugdment
Wunseh,	Wunsch,	entstand in unserem Herzen der **Wunsch**,	better
East	Rast	durch die **Rast** neu gestärkt	better
Übergangspunkt,. das	Übergangspunktr das	ein äußerst lohnender Übergangspunkt, das	equal
großen. Freude	großen, Freude	zu meiner **großen Freude**	equal
halten	hatten	Wir **halten** es nicht mehr aus	worse
là	la	c'est **là** le rôle principal qu'elle joue	worse

Conclusion

- Combination of OCR systems improves quality
- Our decision procedure works in spite of uncorrected training data
- Selection based on unigram frequency is not optimal (*là* $\longrightarrow$ *la*)
  Still untested: higher-order word n-grams
- Ongoing work:
  - Expansion of Text+Berg corpus
  - Further refinement of pattern-based OCR correction and OCR system combination
  - OCR error correction based on morphological analysis

Thank you for your attention!

Questions?